Sig2Lead Ver. 1.0

User Manual

Contents

Sig2Lead Overview 3

Reference 3

Paper Abstract 3

Installation/Configuration of Sig2Lead_v1 4

Introduction 4

Installation/Configuration of RStudio Version 4

A. Installation of R and RStudio 4

B. Download Sig2Lead from Github 4

C. Installation of Dependencies/Libraries 4

Installation of Dockerized Version 5

A. Installation of Docker 5

B. For list of useful docker commands please see: 5

C. Download Sig2Lead Container from Docker Hub 5

D. Open shell and Run Sig2Lead Image Mounting Local Folder 5

E. Open browser and navigate to localhost:3838 6

F. Run app following instructions below 6

Connectivity Analysis Tab 6

A. Define Target Gene Workflow 6

B. Upload a Signature Workflow 8

C. Find Analogs in LINCS Workflow 9

D. Advanced Options 10

Chemical Similarity Analysis Tab 11

A. Chemical Similarity Analysis Options 12

Network Analysis Tab 13

Sig2Lead Overview

Sig2Lead aims to facilitate drug discovery and re-purposing by connecting small molecule and target gene knock-down (KD) transcriptional signatures generated by the LINCS consortium. For a target gene specified by the user, putative inhibitors are identified as those drug-like molecules in LINCS that have signatures concordant with a KD signature of the target. Note that LINCS arguably represents the largest resource for pharmacogenomics to date, with over 20,000 small molecules and about 5,000 gene KDs transcriptionally profiled, thus covering a large subset of the drug-like chemical space and druggable subset of the genome. Furthermore, if a set of candidate molecules is provided, e.g., identified by virtual or experimental screening, these external candidate molecules are ranked based on their chemical similarity to 'concordant' LINCS analogs using a fast chemical similarity search. Additionally, Sig2Lead can be used to prepare input files for docking simulations to be performed in conjunction with connectivity-based analysis t improve the specificity of the search (see paper below), as well as identify LINCS analogs irrespective of their connectivity to a target of interest.

Reference

The Sig2Lead application implements the signature connectivity analysis for drug discovery found in the following paper:

Thorman, A. W., Reigle, J., Chutipongtanate, S., Shamsaei, B., Pilarczyk, M., Fazel-Najafabadi, M., ... & Meller, J. (2020). Accelerating Drug Discovery and Repurposing by Combining Transcriptional Signature Connectivity with Docking. *bioRxiv*.

Paper Abstract

The development of targeted treatment options for precision medicine is hampered by a slow and costly process of drug screening. While small molecule docking simulations are often applied in conjunction with cheminformatic methods to reduce the number of candidate molecules to be tested experimentally, the current approaches suffer from high false positive rates and are computationally expensive. Here, we present a novel *in silico* approach for drug discovery and repurposing, dubbed *connectivity enhanced* Structure Activity Relationship (*ce*SAR) that improves on current methods by combining docking and virtual screening approaches with pharmacogenomics and transcriptional signature connectivity analysis. *ce*SAR builds on the landmark LINCS library of transcriptional signatures of over 20,000 drug-like molecules and ~5,000 gene knock-downs (KDs) to connect small molecules and their potential targets. For a set of candidate molecules and specific target gene, candidate molecules are first ranked by chemical similarity to their 'concordant' LINCS analogs that share signature similarity with a knock-down of the target gene. An efficient method for chemical similarity search, optimized for sparse binary fingerprints of chemical moieties, is used to enable fast searches for large libraries of small molecules. A small subset of candidate compounds identified in the first step is then re-scored by combining signature connectivity with docking simulations. On a set of 20 DUD-E benchmark targets with LINCS KDs, the consensus approach reduces significantly false positive rates, improving the median precision 3-fold over docking methods at the extreme library reduction. We conclude that signature connectivity and docking provide

complementary signals, offering an avenue to improve the accuracy of virtual screening while reducing run times by multiple orders of magnitude.

Installation/Configuration of Sig2Lead_v1

Introduction

Sig2Lead is available in an RStudio version and as a docker container. The RStudio version is available on Github and the dockerized version is available on Docker Hub. The RStudio version requires that R and RStudio be locally installed on the user's computer and requires the installation of all requisite R packages. The dockerized version simply requires a web browser to run the app and does not require local installation/configuration of R, RStudio, or R packages. This manual provides instructions for both versions.

Installation/Configuration of RStudio Version

A. Installation of R and RStudio

The latest version of R is required for configuration of Sig2Lead. At the time of writing this manual, that was version 4.0.3. This version can be downloaded at:

http://www.r-project.org/

Additionally, RStudio is needed and can be downloaded at:

https://www.rstudio.com/products/rstudio/download/

B. Download Sig2Lead from Github

Sig2Lead and associated files can be downloaded from:

https://github.com/sig2lead/sig2lead_v1/

C. Installation of Dependencies/Libraries

Once R and RStudio are installed, shiny must be installed. This can be completed by typing into the R console:

```
install.packages("shiny")
```

All other dependencies and libraries will be installed upon the first time running the application.

This step may not be handled properly on MacOS, requiring a step by step installation of missing libraries. An alternative for Mac users is to use the dockerized version that takes care of all the dependencies.

To run the application, click the  button at the top middle of the RStudio interface.

Installation of Dockerized Version

A. Installation of Docker

Ubuntu: follow the instructions to get Docker CE for Ubuntu.
Mac: follow the instructions to install the Stable verion of Docker CE on Mac.
Windows: follow the instructions to install Docker Toolbox on Windows.

B. For list of useful docker commands please see:

https://www.digitalocean.com/community/tutorials/how-to-remove-docker-images-containers-and-volumes

C. Download Sig2Lead Container from Docker Hub

Go to http://hub.docker.com/repositories/Sig2Lead/ and download Sig2Lead_v1 image

To obtain the docker image and run the container,

[sudo**] docker pull Sig2Lead_v1**

Linux users may need to use sudo to run Docker.

First make sure that port 3838 is free to use for RStudio, (Typically RStudio dockers run on this port, if this port is free ignore the rest of this section). You can stop and kill any othe docker containers on this port by

[sudo] docker stop <container ID> && docker rm <container ID>

To check the container ID run this command:

docker ps -a

D. Open shell and Run Sig2Lead Image Mounting Local Folder

To start an RStudio session, open a browser and type in the address bar http://localhost:3838 on Mac or Linux systems when 3838 port is used. Use RStudio for the user name and password.
Host URL on Ubuntu and Mac is localhost, if accessed locally. On Windows, the IP is shown when Docker is launched by double-clicking the Docker Quickstart Terminal icon on desktop, or it can be obtained from the output of docker-machine ls in the interactive shell window.

Open a command prompt and run following command:

docker run -*v* <*local_path*>:/srv/shiny-server/userfile -*d* Sig2Lead_v1

The <local path> is the path of the local directory that contains the smiles or sdf file of your added compounds.

For Windows, the following line is an example with local directory "C:\Added_Compounds"

docker run -v C:\Added_Compounds:/srv/shiny-server/userfile -d Sig2Lead_v1

For MAC/UNIX, the following line is an example with local directory "/User/Added_Compounds"

docker run -v /User/Added_Compounds:/srv/shiny-server/userfile -d Sig2Lead_v1

E. Open browser and navigate to localhost:3838

Open browser and navigate to the following url: http://127.0.0.1:3838

F. Run app following instructions below

Follow the instructions below for running the app.

Connectivity Analysis Tab

In the next several sections, several distinct workflows organized in 3 different tabs are discussed and illustrated using use cases. The first of those sections below describes the main tab where the primary workflow can be initiated by defining the target gene.

A. Define Target Gene Workflow

This is the standard workflow for identifying small molecule inhibitors or activators of a target of interest. Within this workflow, a gene of interest is required and optionally, a user-defined list of compounds for scoring can be provided in SDF or SMILES format. Sig2Lead collects all data from genetic knockdowns within LINCS and identifies compounds that generate highly concordant transcriptional signatures to these genetic knockdowns, i.e., putative target/pathway inhibitors. If an external set of candidate molecules is provided, e.g., identified by virtual or experimental screening, these user-provided candidate molecules are ranked based on their chemical similarity to 'concordant' LINCS analogs using a fast chemical similarity search. This method allows scoring of small molecules to be tested for the purpose of library reduction. Results of added compounds will be ranked in descending order of similarity to LINCS analogs, with ties broken by concordance scores in the table titled "My Candidates Ranked" and compounds identified from within the LINCS library will be scored only by their concordance values to a knockdown of the target gene of interest in the table titled "LINCS Compounds Ranked" (**Error! Reference source not found.**). These tables can be downloaded using the download buttons just below the "Go!" button.

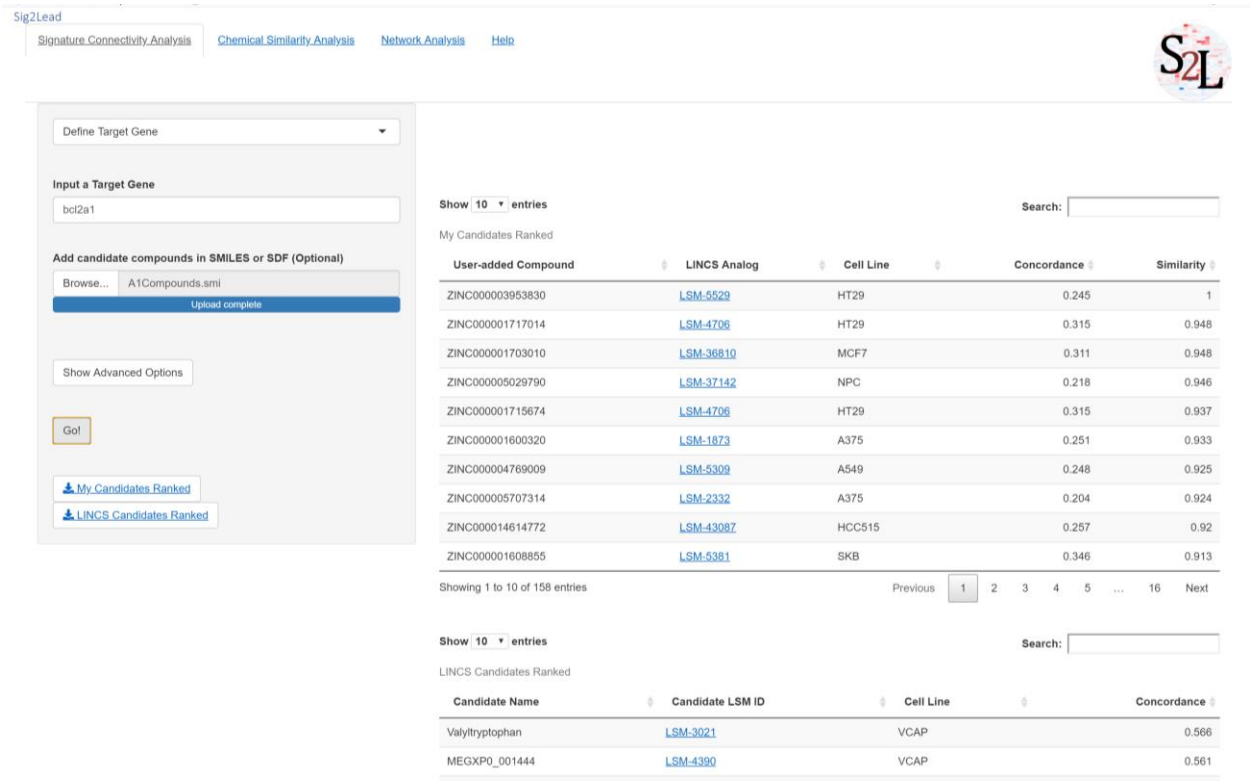

Figure 1: Define Target Gene workflow. In this primary workflow, the user defines a gene of interest to identify putative pathway/target inhibitors. If the user defines a set of compounds from some other analyses, those compounds will be compared to the LINCS library for analogous compounds and scored by similarity and concordance in the upper table (My Candidates Ranked), which will only appear in the event of added compounds. The lower table (LINCS Candidates Ranked) will always appear when running this workflow and scores all LINCS compounds with a concordance above the threshold in descending order of concordance.

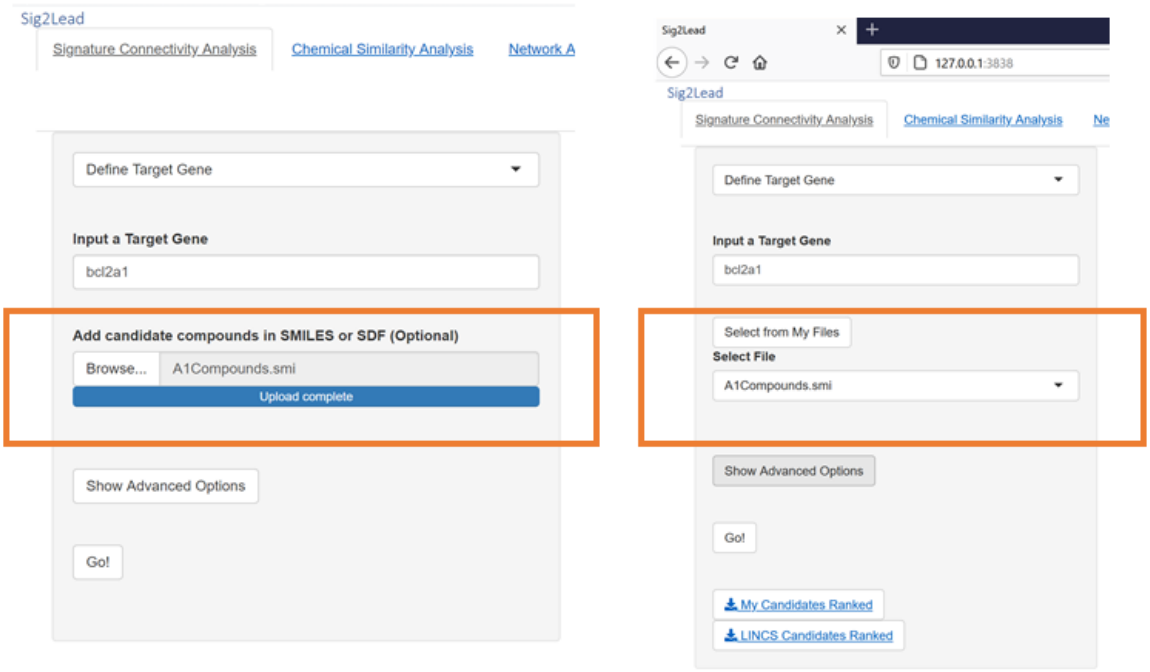

Figure 2: The two file input controls for the RStudio version (left) and the dockerized version (right) are shown in the above figure. The RStudio version (left) has a standard upload control in which the user can press the "Browse" button and navigate to the directory with the ".smi" or ".sdf "files of user-provided compounds. The file input control for the dockerized version (right) is slightly different with a button labeled "Select from My Files" which when pressed lists ".smi",".txt", and ".sdf" files in the user's local mounted directory (instructions for mounting a directory when launching the dockerized version are described in the 'Installation of Dockerized Version' section) which the user selects.

B. Upload a Signature Workflow

This workflow allows users to upload a signature of their own to search for potential inhibitors to a target gene unavailable on LINCS, or molecules that generate a similar signature to some other system perturbation that is otherwise undefined. In this workflow, users define a signature using one of the formats defined at (www.ilincs.org/ilincs/signatures/main/). This pipeline otherwise follows the same pipeline as the "Define Target Gene" workflow.

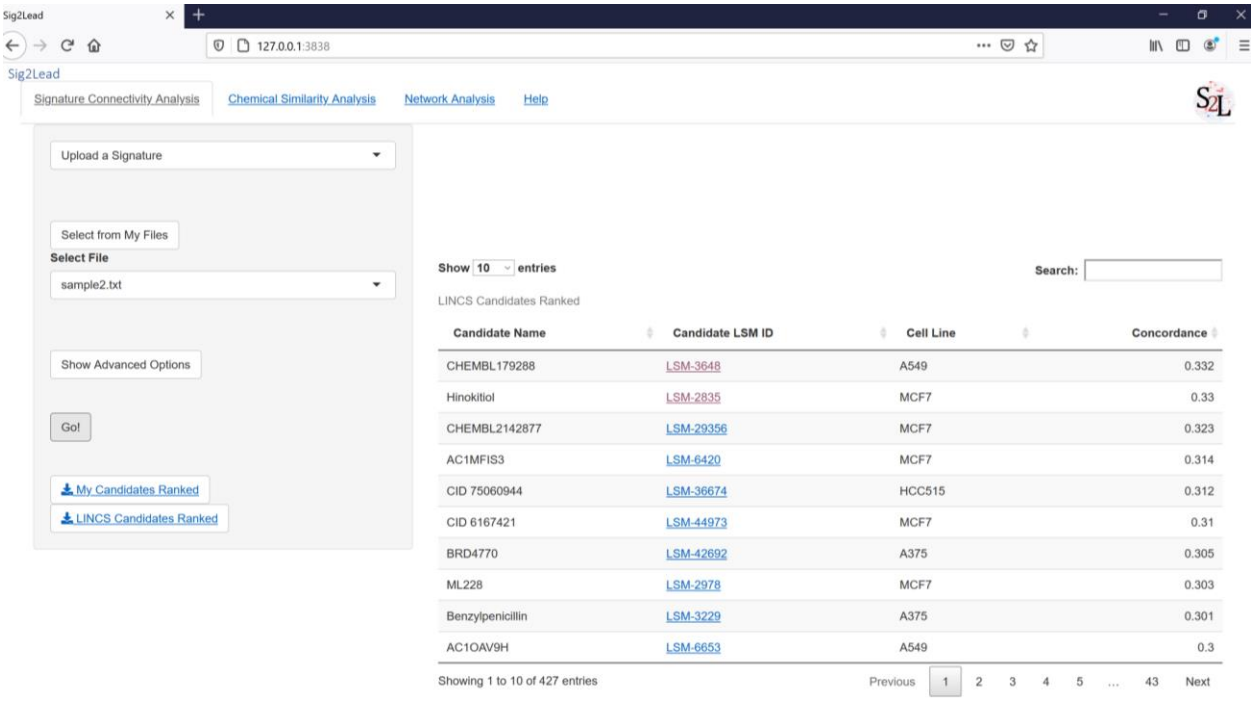

Figure 3: Upload a Signature Workflow. This workflow allows one to query iLINCS (see ilincs.org) to find small molecules in LINCS that have signatures concordant with a user-provided transcriptional signature. The signature file should be saved as a tab-separated text file with gene symbol, log of differential expression value, and p-value comprising the columns in that order. The output of this workflow is a table of concordant LINCS small molecules to the uploaded signature.

C. Find Analogs in LINCS Workflow

This workflow simply identifies LINCS analogs to user-defined compounds, irrespective of their connectivity to a target. This requires added compounds and is useful in determining if there are transcriptionally profiled analogs in the LINCS library. LINCS compounds are all drug or drug-like, so this may be a simple filter to remove compounds that do not contain normal drug-like structures. Additionally, it was used in benchmarking as a baseline similarity of various compound libraries to the LINCS small molecules.

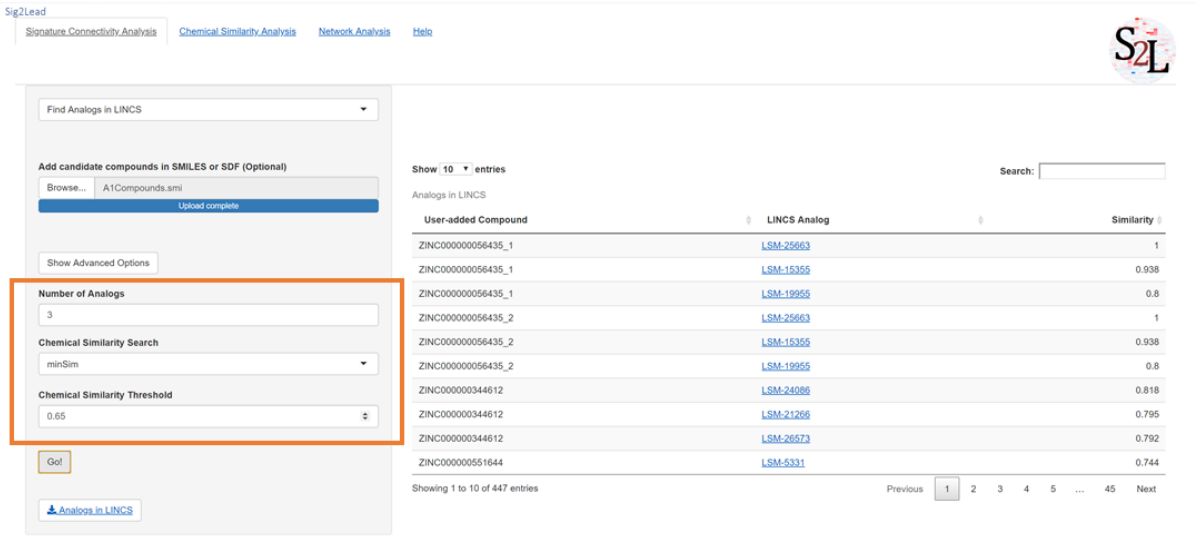

Figure 4: Find Analogs in LINCS Workflow. This workflow allows the user to find small molecules included in LINCS library that are structurally similar to user-provided compounds. The Advanced Options to change the number of LINCS analogs to be returned for each user-provided compound bounded by the chemical similarity threshold (also adjustable). The user can also use either the ultrafast minSim algorithm or slower fpSim algorithm to compute chemical similarity.

D. Advanced Options

A. Users can select to identify activators instead of inhibitors. Functionally, this is searching for molecules that are discordant to the gene knockdown of interest instead of concordant.
B. Users can change the method of similarity search. The default, minSim is an exact fast chemical similarity search utilized by Sig2Lead. This is a novel contribution within Sig2Lead that is currently unpublished as its own method. For comparison, an established fpSim function to compute chemical similarity is available as a slow option (it typically runs about 100x slower than minSim).
C. Finally, users can change the concordance threshold. The default of 0.2 is the minimum allowed threshold and corresponds to the threshold used in benchmarking for the ceSAR publication. For some targets, a more stringent cutoff may be required to increase the specificity. The range of these concordances can theoretically be from -1 to 1, but realistically below 0.2 have no significant concordance and very few generate a concordance above 0.5 or 0.6 (see the Sig2Lead/ceSAR manuscript).

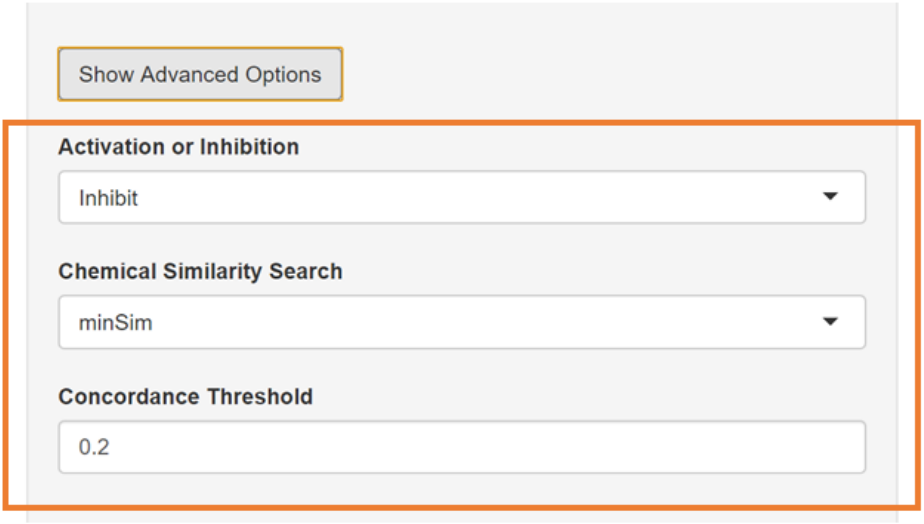

Figure 7. Define Target Gene Workflow Advanced Options.

Chemical Similarity Analysis Tab

After searching LINCS for analogs, compounds can be further analyzed through chemical similarity using the "Chemical Similarity Analysis" tab. This would be done as a standard if only Sig2Lead is run for compound identification but is not necessary to run a ceSAR analysis as described in the linked publication. To run chemical similarity clustering click the "Run SAR" button. This will initiate a clustering analysis that compares concordant compounds from LINCS for the target and user-added compounds and clusters them by chemical similarity to one another. By default, 5,000 compounds will be analyzed (this can be changed under advanced options if an extended time is acceptable). The output will be in the form of two figures, a heatmap and an MDS plot, and a table of centroids for each cluster (**Error! Reference source not found.**). The heatmap is generated through hierarchical clustering and shows a distance matrix comparing each compound identified through LINCS (Green) or added by the user (Blue). The MDS plot is an alternative view of the hierarchical clustering showing relative distances between clusters of compounds. The radius of each pie chart corresponds to the size of the cluster.

Figure 5. Example of chemical similarity analysis. SAR type analyses can be performed for either 'concordant' LINS compounds (here shown in Green), or user provided candidate molecules (here in Blue), or together to identify distinct classes of chemical moieties within the set of user provided compounds and their 'concordant' LINCS analogs. The heatmap in the left panel shows pairwise chemical similarity pattern with the compounds analyzed as both columns and rows, i.e., the diagonal represents identity (Tanimoto coefficient of 1.0). Note that user defined candidates indicated by blue ticks in the top bar are scattered throughout, i.e., they belong to several distinct classes of compounds with most of them in the middle 'big' cluster indicated by a large rectangular block of high similarity scores in the middle of the heatmap. This is further highlighted in the right panel that shows individual clusters identified using the MDS 2-dimenshional projection of pairwise similarities, with the 'big' circle in the bottom left corner corresponding to the central mixed cluster in the heatmap.

A. Chemical Similarity Analysis Options

A. For clustering, different Tanimoto similarity thresholds can be selected. The default is 0.75, which groups compounds of modest difference together, but can be adjusted to make more (closer to 1) or less stringent clusters (closer to 0). This will only be reflected in the MDS plot and the table of representatives.
B. The minimum cluster size for inclusion in the MDS plot can also be changed. By default, compounds that are not structurally related to at least two others at the specified threshold are independent and a representative of that group cannot be retrieved. The user can instead

cluster all compounds regardless of cluster size or increase this threshold to only consider those from large clusters of related compounds.

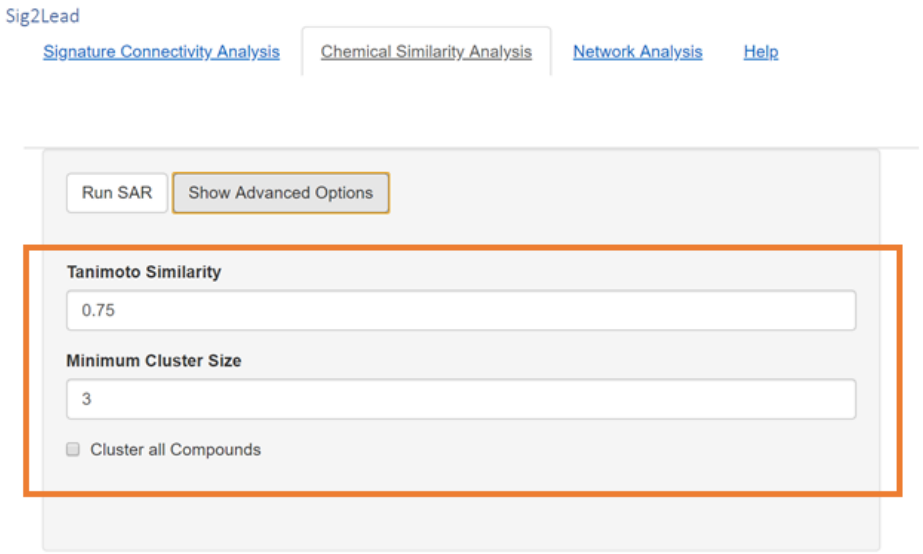

Figure 8. Chemical Similarity Analysis Advanced Options.

Network Analysis Tab

Finally, Network connectivity analyses can be performed after chemical similarity analyses using STITCH on the “Network Analysis” tab. This analysis is intended to scrape any known information about identified compounds and their interactions with members of the pathway of interest. This step can be performed either through a global view (all identified compounds) or on a cluster by cluster basis (much faster). When running Global STITCH analysis, only the compounds found in clusters of sufficient size, as determined on the “Chemical Similarity Analysis” tab’s advanced options will be included unless the “Shows all clusters” box is checked. The Global STITCH analysis can be slow, depending on the number of compounds added.

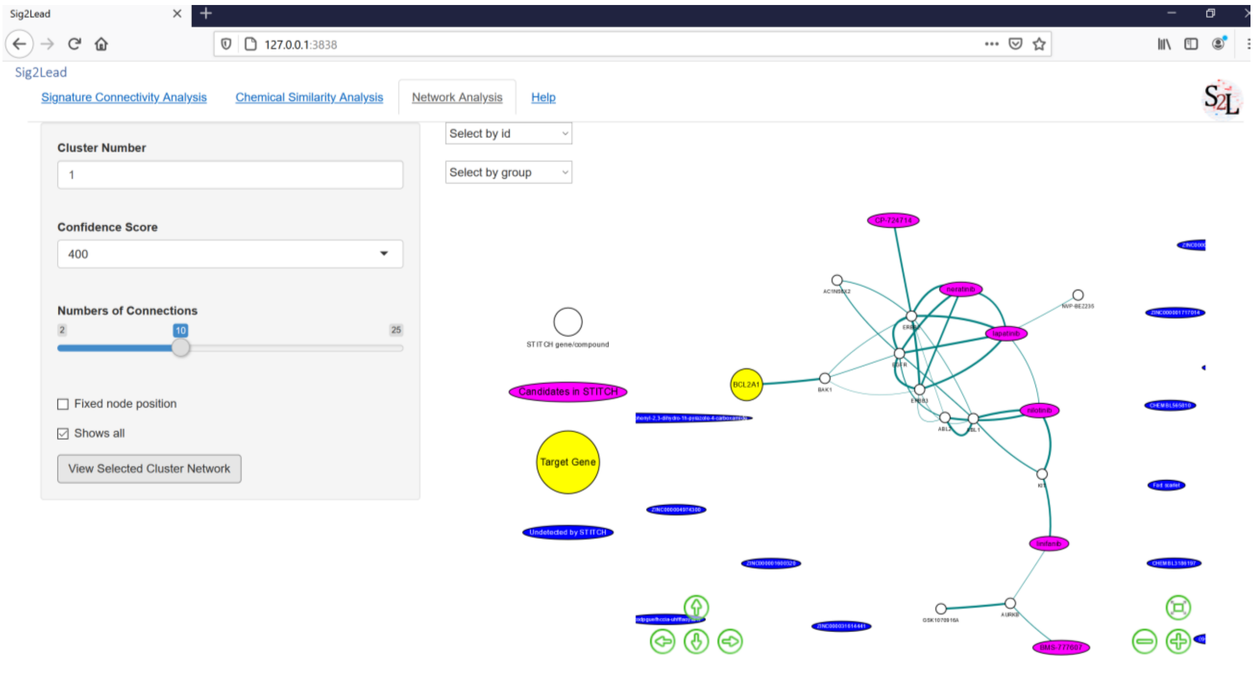

Figure 6: Cluster Network STITCH. Under the STITCH Network analysis, known interactions will be mapped between the target gene and putative compounds submitted via Sig2Lead. The target gene is shown in yellow, other STITCH derived genes within the pathway are shown in white, and compounds identified via Sig2Lead are shown in magenta. The figure shows additional compounds to be considered as putative pathway inhibitors (blue) based on their chemical similarity to known pathway or target inhibitors (magenta).